# Modelo determinista para el análisis sintáctico: una perspectiva algorítmica y computacional de la lingüística

*Julio Meroño Sáez*

**Resumen**

Spanish is the fourth most spoken language across the globe, right behind English, Mandarin and Hindi. Thus, it is of high interest to have an accepted linguistic standardization, something that has historically been carried on by numerous entities, and that, nowadays, is done by the Asociación de Academias de la Lengua Española (Association of Academies of the Spanish Language). The works that have been published in this context, whose sole purpose is the aforementioned standardization of the Spanish language, fail to define with the required rigor the relations and hierarchies observed in Spanish, being limited, instead, to providing criteria for their distinction and classification. Even if unnecessary in numerous other fields, such a standardization closes the gap to achieving noteworthy developments in fields such as education, legislation and the construction of artificial intelligence models. Therefore, the main goal is to provide a model that contains these characteristics which, additionally, is mathematically compatible with an implementation to be done afterward as a computer program, with intentions to later provide a means of access to the general public through a web interface. To achieve such a goal, throughout this paper, the proposed model is split into several inner levels and the algorithms, data structures and formats in use are thoroughly explained.

El español es la cuarta lengua más hablada globalmente, detrás del inglés, el chino mandarín y el hindi. Por ello, es fundamental que exista una estandarización lingüística, cometido realizado por numerosas entidades a lo largo de la historia del español, y actualmente llevado a cabo por la Asociación de Academias de la Lengua Española. Sin embargo, las obras publicadas y reconocidas cuyo propósito es esta misma estandarización de la lengua no consiguen definir, con el nivel de

rigor suficiente, las relaciones y jerarquías que se observan en la lengua, sino que se limitan a proporcionar criterios para su distinción y clasificación. Aun innecesario para numerosos ámbitos, este rigor proporcionaría los medios necesarios para alcanzar considerables avances en ámbitos como la educación, la legislación y la construcción de inteligencias artificiales. Es por ello por lo que se pretende conseguir un modelo de estas características, que, además, sea matemáticamente compatible con una posterior implementación en forma de programa informático y disposición al público mediante una interfaz en línea. Para ello, en el trabajo de investigación, se disecciona el modelo adaptado propuesto en niveles a lo largo de su construcción y, posteriormente, se explican con detalle los algoritmos, estructuras de datos y formatos que han sido de relevancia y utilidad.

# Índice

# Parte I. Prólogo

## 1. Introducción

El español o castellano es una lengua iberorromance y es la lengua nativa de más de 450 millones de hablantes en el mundo, a lo que se le suman los 75 millones de hablantes no nativos alrededor del planeta. Esto la sitúa como la cuarta lengua más hablada globalmente, detrás del inglés, el chino mandarín y el hindi. Es por ello por lo que la existencia de una lingüística estandarizada es fundamental, al permitir la comunicación internacional sin cabida a ningún tipo de ambigüedad y al mantener la integridad de la lengua. La Asociación de Academias de la Lengua Española, junto a sus correspondientes academias nacionales (entre ellas, la Real Academia Española), son las encargadas de dicho cometido, proporcionando una estandarización de la gramática, morfología, léxico y fonética del español alrededor del mundo hispanohablante.

Sin duda, la gramática es tan fundamental como las otras ramas de estudio de la lingüística. Esta es la encargada de estudiar la organización y la función de las palabras dentro de una oración, junto a la agrupación de palabras en unidades mayores con significado y función, denominadas sintagmas. A su vez, dichos sintagmas se organizan por categorías sintagmáticas que engloban y marcan su función general. El estudio y el estandarizado de la gramática en el español se remonta a finales del siglo XV, con autores como Antonio de Nebrija, llegando y continuando hasta la actualidad con teorías y propuestas transversales a la mayoría de lenguas, como las propuestas por Noam Chomsky, o trabajos más enfocados en el español.

Tradicionalmente, y aún en épocas recientes, se ha realizado un estudio sintáctico más específico para cada categoría sintagmática, donde los sintagmas (o grupos, como se hacen llamar en ciertos modelos) poseen estructuras fundamentalmente distintas entre tipos, y donde se da lugar a numerosos casos límite a tener en cuenta. Aún así, dicho análisis ha resultado ser considerablemente más efectivo y simple a la hora de su uso superficial por lingüistas procedentes de otras ramas, o en el ámbito educativo como medio de enseñanza. Sin embargo, desde finales del siglo anterior y gracias a la teoría estándar propuesta por

Noam Chomsky, se ha dado lugar a otro enfoque distinto a la gramática, denominado gramática generativa. En la mayoría de teorías y modelos asociados con dicha gramática, se formula que todas las agrupaciones a las que se les denomina sintagmas comparten una estructura fundamental, que aun pudiendo variar en el tipo de unidades y tipos de sintagmas agrupados para formarlo, siempre se estructuran de la misma manera, sin casi ninguna cabida a variación.

El estudio de la gramática en el español, especialmente de la sintaxis, ha resultado en el creciente interés por el uso de las nuevas tecnologías y de la capacidad computacional de este siglo para el análisis sintáctico de frases, sintagmas individuales y oraciones. Por desgracia, el uso en su mayoría de la gramática tradicional para la creación de modelos y algoritmos que cumplan dicha tarea dificulta considerablemente la detección y el correcto análisis de ciertos casos especiales o límite que se pueden dar, o imposibilita casi por completo el proporcionamiento de más información en cuanto a la función que realizan ciertos sintagmas dentro de agrupaciones mayores, factor que la gramática tradicional no suele tener en cuenta.

Es por ello por lo que este trabajo de investigación propone un análisis determinista en forma de modelo adaptado de la gramática del español, basándose fundamentalmente en la gramática generativista aportada en la obra Sintaxis generativa del español [Zagona, 2006], pero incluyendo modificaciones tanto propias como de otros autores que permitan una mayor facilidad para el procesado de oraciones mediante implementaciones computacionales de dicho modelo. Además, se pretende diseñar y llevar a cabo una implementación tipo del modelo formulado previamente que sea capaz de poder utilizarse en ordenadores convencionales sin requerir de un excesivo coste computacional, para su posible y posterior publicación mediante una interfaz en línea y de acceso libre para su uso público.

## 2. Justificación

Son decenas de millones los estudiantes que se encuentran actualmente aprendiendo español en las escuelas. En específico, contando únicamente los estudiantes extranjeros y no procedentes de ningún país hispanohablante, se estima que el número se encuentra en el orden de los 24 millones [Instituto Cervantes, 2022]. Entre toda la materia y contenidos aportados en la clase a dichos estudiantes durante sus horas lectivas, una gran y significativa parte consiste en el estudio de la propia lengua. Además, la utilización del estudio y análisis sintáctico mediante separación en sintagmas o grupos encuadrados se ha vuelto a lo largo de los años el estándar a nivel nacional y, en ciertos casos, internacional. Aunque los estudios que lo afirmen son relativamente escasos, predominantemente la experiencia ha demostrado a las instituciones que planifican las competencias a tener en cuenta a la hora de enseñar en las aulas que el entendimiento del funcionamiento de la lengua implica un mejor uso de esta, tanto social y oral como formal y escrito, y es fundamental para la preservación del español, razón junto a las anteriores por las que se muestra un interés significativo en su enseñanza en los centros educativos a lo largo de todo el territorio español. Además, se ha de tener en cuenta la reciente y creciente adopción de esta metodología en otros países de habla española, lo que implica una mayor impulsión del conocimiento interno del funcionamiento de la sintaxis española.

Si bien, el uso de la gramática tradicional como herramienta didáctica se remonta a, aproximadamente, el siglo anterior, la cantidad de material de ejercicios para los estudiantes es particularmente escaso, especialmente cuando se pretende comparar con otras materias donde no solo existe mucho más material, sino que es posible su generación procedural y automática. En principio, lo comentado no debería tener ninguna implicación negativa, pues al tratarse de un estudio de la gramática española en forma de sintaxis basada en teorías tradicionales, la generación de ejercicios resulta tan simple como obtener, buscar o producir por la propia cuenta del profesorado oraciones que se adapten al nivel de los estudiantes. Sin embargo, se ha de considerar y tener en cuenta que, para comprobar la correctitud y validez de las respuestas otorgadas por los estudiantes a espera de ser evaluadas, resulta tedioso tener que realizar una comprobación mental lenta y laboriosa

de dichos resultados, que puede además dar lugar a numerosos errores de comprobación, no solo beneficiosos sino también perjudiciales en ciertos casos para el alumnado. Por otro lado, desde el punto de vista estudiantil, la comprobación independiente de resultados es completamente imposible, pues se carece de un punto de referencia objetivo para comparar, por lo que estos carecen de una forma de poder practicar de manera efectiva e ilimitada, sin depender del continuo trabajo del profesorado.

Lo que diferencia principalmente a las personas de cualquiera de los dos puntos de vista anteriores con los ordenadores y dispositivos electrónicos es la capacidad de poder tomar decisiones aplicando el entendimiento horizontal, subjetivo y basado en la experiencia y no en reglas a aplicar. Por lo tanto, la tarea de producir análisis sintácticos utilizables en el ámbito de la creación de asistentes basados en la inteligencia artificial se considera un tema popular y continuamente en desarrollo y debate en investigaciones de las últimas décadas.

Este último caso nos lleva al ámbito judicial, donde se muestra una carencia de estandarización intuitiva del significado de los textos legislativos, los contratos y los acuerdos formales, dando lugar a que existan profesiones cuyo único propósito sea el de garantizar la ausencia de vacíos legales a la hora de producir dichos papeles. Todo esto ocurre por una vaga especificación abierta a interpretaciones de la gramática utilizada en este campo.

La motivación de este trabajo de investigación es, entonces, aportar un pequeño pero significante grano de arena a la base de lo que será el mundo en los años y las décadas por venir, tanto en el ámbito de la educación, como en el de la tecnología, como en el judicial, como en cualquier otro que se pueda plantear (lenguaje matemático, diplomacia, etcétera). Para ello, se plantean diversos objetivos escalonados que se formulan en el apartado a continuación.

# 3. Objetivos del trabajo

El objetivo principal de este trabajo de investigación es, en resumidas cuentas, poder aportar un modelo descriptivo y determinista pero empírico de la gramática del español, basándose en una agrupación de teorías y definiciones de la gramática de autores anteriores y antecedentes, mezclado con un enfoque propio matizado por la mentalidad computacional y matemática, que pueda utilizarse como herramienta de formalización, educación y comprobación, pero además como punto de partida para futuras investigaciones.

Por otro lado, este objetivo ya mencionado da lugar a ciertos enfoques alternativos pero compatibles por los que llevar la investigación, en forma de objetivos secundarios:

1. Crear una herramienta de uso fácil e inmediato para el análisis de oraciones y sintagmas en español, disponible de forma completamente gratuita al público, para su uso por profesorado, estudiantes, políticos, diplomáticos, abogados, etcétera.

2. Aportar una estructura de modelo fácilmente implementable que permita su uso de forma más eficaz a la hora de realizar análisis computacionales de la sintaxis del español, no solo como punto de apoyo para el anterior objetivo sino también como base para modelos de asistencia mediante inteligencia artificial, chatbots y cualquier tecnología similar.

3. Agrupar los esfuerzos provenientes de las distintas corrientes lingüísticas existentes para propulsar la gramática española en un trabajo recopilatorio que, además, introduce ciertas variaciones partiendo de un enfoque más tecnológico, algorítmico y modélico.

# Parte II. Modelo adaptado de la gramática española

En esta parte del trabajo de investigación, se diseña y redacta un modelo adaptado de la gramática del español, con la intención de su posterior uso en la implementación del mismo mediante técnicas propias del ámbito de la programación y la algoritmia. Previamente a ello, sin embargo, se nombra la base a partir de la que se construye el modelo, y se menciona en detalle la estructura, metodología de diseño y paradigma utilizado para la creación del mismo.

Es de relevancia para el entendimiento del trabajo tener en cuenta que, en la redacción del modelo, se parte de una base nula, posiblemente reasignando definiciones incompatibles a cierta terminología en uso con el fin de simplificar la producción y aumentar la claridad de este. Además, las propuestas de clasificación, análisis y estructurado pueden poseer disparidades significativas con las procedentes de las corrientes gramaticales actuales, incluso, en definiciones generalmente aceptadas previamente. Esto es debido, por otro lado, a los intentos de acercar el modelo a una redacción lo suficientemente formal en el sentido algorítmico como para su utilización sin posible lugar a dudas en las implementaciones.

Por último, por limitaciones de tiempo, de personal y de antecedentes previos a este trabajo de investigación, resulta considerablemente dificil, incluso imposible en la práctica, tratar de definir la totalidad de la gramática española en el mismo, por lo que se ha decidido reducir la lengua en cuestión a un subconjunto del español. Con esto se pretende decir, adicionalmente, que cualquier enunciado válido en esta versión reducida del español es, también, un enunciado válido en el español real. También se descartan ciertas variaciones internas dentro las estructuras sintagmáticas, limitándolas a formas y ordenamientos fácilmente procesables y comunes en el español, mencionados posteriormente en el trabajo.

## 4. Fuentes e investigaciones previas

El modelo presentado y definido a lo largo de las siguientes secciones obtiene su inspiración y punto de partida en diversas corrientes gramaticales y lingüísticas existentes en el momento. Sin embargo, dichas corrientes carecen del rigor matemático o algorítmico necesario para la realización de implementaciones formalmente verificables. De esta manera, el marco bibliográfico constituido por trabajos de relevancia para el diseño del modelo se encuentra severamente limitado. Esto conlleva, como se ha mencionado previamente, la necesidad de simplificar y reducir la gramática tratada en el modelo a un conjunto reducido y delimitado por lo fácilmente deducible y extrapolable a definiciones formales de lo explicado en las fuentes bibliográficas e investigaciones previas de este trabajo de investigación. Sin embargo, esto implica que las fuentes en uso juegan un papel meramente inspiracional o de guía en las definiciones aportadas, por lo que se ruega discreción a la hora de su interpretación en las secciones posteriores.

A continuación, se mencionan brevemente las fuentes que se han utilizado para la construcción y el diseño del modelo:

1. *Corpus AnCora.* [Antònia Martí et al., 2007] Este corpus español-catalán posee diversas anotaciones morfológicas, sintácticas y semánticas de significativa utilidad para este trabajo de investigación. En especial, los dos lexicones extraídos del mismo, AnCora-Verb y AnCora-Nom, forman la base del procesado semántico efectuado en la implementación del modelo. Las agrupaciones en argumentos de los roles semánticos presentes en el modelo se realizan con la intención de simplificar la tarea de llevar su análisis a la práctica, alineándolas con las presentes en estos lexicones.

2. *Nueva gramática de la lengua española.* [Asociación de Academias de la Lengua Española [ASALE], Re
   Desde su publicación en 2009, esta obra académica ha servido como actualización de la gramática normalizada por las mismas entidades en la primera mitad del siglo anterior. En este trabajo de investigación, la gramática procedente de la misma ha constituido la base para gran parte de la terminología y definiciones en uso y la clasificación de elementos, especialmente en los apartados más relacionados con la morfología y la enumeración de rasgos gramaticales.

3. *Corpus del español del siglo XXI.* [Real Academia Española [RAE], 2023] Las tablas de frecuencias de este corpus han sido de considerable utilidad para la formación de la base de datos de elementos gramaticales en uso en la posterior implementación, además de proporcionar un análisis de rasgos gramaticales significativamente exhaustivo, razón por la que esta fuente ha sido crucial para el trabajo de investigación. Es por ello por lo que, para el diseño del modelo, se ha tomado la leyenda de rasgos gramaticales proporcionados en el análisis como punto de partida para su nombramiento y clasificación.

4. *Sintaxis generativa del español.* [Zagona, 2006] En este libro, se proporciona una visión global de la sintaxis del español desde un punto de vista basado en la gramática generativista originaria de Noam Chomsky. De esta visión surge gran parte de las estructuras de las cláusulas y los sintagmas tratadas en este trabajo de investigación, además de, en general, servir como inspiración movida por las corrientes generativistas para la planificación y el diseño de ciertas secciones del modelo, tanto en los apartados enfocados en la morfología como en los que tratan principalmente la sintaxis del subconjunto del español en cuestión.

# 5. Estructura y visión del modelo

Dentro del modelo redactado a continuación, se proporciona un enfoque mayoritariamente funcional y empírico, evitando comprometer la simplicidad del modelo que se perdería al pretender tratar temas como el origen natural de ciertas estructuras. Por ello, se proporcionan declaraciones que repliquen las construcciones naturales de la lengua con la mayor precisión posible dentro de los límites prácticos del trabajo.

La estructura del mismo consiste en tres niveles fundamentales, ordenados según su dependencia en los niveles anteriores y según el grado de abstracción presente en cada uno de ellos. De esta manera, se comienza tratando el nivel léxico-morfológico, donde se trata la base léxica de la que se parte, sus rasgos gramaticales y la clasificación de estos. A continuación, se menciona el nivel sintáctico, en el que se explican las distintas estructuras válidas para la agrupación de lexías en enunciados o subdivisiones de los mismos. Por último, se describe el nivel semántico, cuya función es la de extraer y procesar el significado de las estructuras formadas mediante la asignación de roles.

De cualquier manera que se plantease el trabajo, sin embargo, siempre sería necesario partir de una base de conocimientos previos no tratados en el mismo, por lo que, durante la creación del modelo, ciertos apartados han sido omitidos por cuestiones de limitaciones o de conveniencia. Específicamente, dos casos son inmediatamente perceptibles y de relevancia:

1. El nivel léxico-morfológico presenta un vacío de contenido en cuanto al origen, la formación y la flexión de elementos gramaticales, pues, al partir del análisis exhaustivo de elementos previamente realizado por la Real Academia Española, resulta totalmente innecesario definir el proceso a llevar a cabo para efectuar dicho análisis. Así, se ignora en el trabajo de investigación la relación existente entre la organización de los morfemas en los elementos y los rasgos gramaticales que surgen como producto de esta organización.

2. El nivel semántico utiliza una agrupación de roles en argumentos que, aun lógica y válida a la vista, resulta aparentemente arbitraria en cuanto a su selección.

Dicha agrupación proviene, sin embargo, de un intento de compatibilizar los argumentos definidos con los presentes en los lexicones procedentes del corpus *AnCora*, posteriormente utilizados en la implementación de este nivel del modelo.

También cabe destacar que la omisión de análisis de unidades mayores que el enunciado se hace con la intención de simplificar el modelo para su factibilidad, pues el análisis de estructuras textuales requiere de un conocimiento preciso sobre su formación y de una cuantización analítica de las estructuras posibles, cosa que escapa el alcance del trabajo de investigación y requeriría de recursos mayores que los disponibles en el momento.

Por último, a lo largo de toda la redacción que abarca el modelo a continuación, se utiliza en las secciones una notación específica para definir ciertos apartados o estructuras formalmente. Para evitar cualquier tipo de confusión, y definirla en su totalidad en los casos donde esta difiera de la norma, se mostrarán explicaciones completas y exactas, previas a su utilización en el trabajo.

# 6. Nivel léxico-morfológico

El nivel léxico-morfológico es el primer nivel a tratar en el modelo, pues trata la base léxica sobre la que se construyen, posteriormente, las estructuras internas de los enunciados. Además, se describen todos los posibles rasgos de las unidades gramaticales, con el fin de poder utilizarlos para hacer posible su análisis en el nivel sintáctico y en el nivel semántico. Para ello, se introduce a *la lexía* como unidad elemental sobre la que construir el resto del modelo.

La lexía es un elemento gramatical cuyas partes presentan un elevado índice de inseparabilidad morfológica o forman una función distinta, no necesariamente presentando cierto valor semántico, que cualquier subconjunto de dichas partes. Las lexías engloban términos más convencionales como la palabra y les aporta una definición menos problemática, pero no se limitan únicamente a estos.

Según su estructura morfológica y visual, las lexías se pueden clasificar de la siguiente manera:

1. *Lexías simples.* Constan de un conjunto de morfemas carente de espaciado o separación simbólica alguna. Se incluyen aquí tanto formas simples como compuestas de las palabras, a diferencia de la clasificación convencional de lexía, pues permite un procesado más simple de las mismas.

2. *Lexías complejas.* También denominadas como locuciones, se componen de varios morfemas donde sí se presenta una separación mediante espacios o símbolos. Cada isla de morfemas se podría considerar equivalente a una lexía simple, sin embargo, la totalidad de la lexía aporta una función y un valor semántico distintos a cada una de las lexías simples discernibles en su interior, o cualquier agrupación de estas.

3. *Lexías textuales.* Los nombres propios, los términos numéricos, las expresiones hechas, las citaciones y cualquier término insondable por las herramientas del ámbito gramatical se consideran lexías textuales, pues difieren de las lexías simples y las lexías complejas en que su significado proviene de una asignación artificial del mismo, de

un contexto externo al entorno o de una interpretación poética de una o más lexías de otra clase.

4. *Símbolos.* Estos son constituidos por cualquier marcación indivisible pero estrictamente individual con una función exclusivamente gramatical. Sus usos incluyen, pero no se limitan a, la delimitación o la coordinación de elementos gramaticales, y no constan en ningún alfabeto.

Por otro lado, según su función y la presencia de valor semántico, se pueden clasificar como se presenta a continuación, siendo esta la clasificación de mayor utilidad a lo largo del modelo:

1. *Lexías semánticas.* Se diferencian al poseer un valor semántico y al constituir casi la totalidad del repertorio léxico. Los adjetivos, los adverbios, los sustantivos y los verbos son los cuatro tipos de lexías semánticas.

2. *Lexías determinantes.* Tienen como función la de concretar o determinar lexías, tornando conceptos en referencias. Además, pueden aportar rasgos gramaticales adicionales a dichas lexías, o reforzar los rasgos ya presentes en ellas.

3. *Lexías gramaticales.* Estas cohesionan diversas lexías entre sí, formando jerarquías y enumeraciones, o sirven para cambiar la función de otras lexías. Se incluyen aquí las conjunciones, las preposiciones y los símbolos.

## 6.1. Rasgos gramaticales

*Los rasgos gramaticales* son propiedades inherentes de las lexías que pueden tomar valores cuantizados y limitados que se encuentran definidos según y para cada rasgo presente. Su presencia se encuentra mayoritariamente condicionada por la clase de lexía en cuestión y, en ciertos casos, por la subclase de la lexía. La utilización de rasgos gramaticales para la organización de lexías y la toma de decisiones analíticas durante su procesado permite la simplificación de ambos procesos, además de su formalización matemática. Estos

se nombran de la siguiente manera:

$r$ = [RASGO : VALOR]; $R = \{r_1, r_2, \cdots, r_n\}$

$r \sim s \iff r$ y $s$ representan el mismo rasgo (son *homónimos*)

$r \equiv s \iff r$ y $s$ representan el mismo rasgo, con el mismo valor (son *equivalentes*)

Entre la selección de rasgos gramaticales que se pueden observar como parte de las lexías, algunos de ellos se muestran con una significativamente mayor frecuencia, incluso, en ciertos casos, a lo largo de diversas clases y subclases de lexías. Estos, denominados *rasgos trasversales*, muestran que, en el español, existen propiedades intrínsecas de las lexías meramente por definición de las mismas. La lista completa de tales rasgos es la siguiente:

- [ANIMADO : SÍ/NO]. Marca si el objeto referenciado posee la capacidad de actuar como agente de un enunciado, donde sí está animado, o si las funciones que puede realizar la referencia excluyen la función de agente, donde no está animado.
- [DETERMINADO : SÍ/NO]. Especifica si se referencia a una instancia específica de un objeto, donde *sí* está determinado, o si se referencia al concepto del mismo, donde *no* está determinado.
- [FUNCIÓN : AGENTE/CIRCUNSTANCIAL/EVENTO/MODIFICADOR/OBJETO/PACIENTE]. Limita el rango de funciones que una lexía puede realizar dentro de una cláusula, variando entre *agente*, *circunstancial*, *evento*, *modificador*, *objeto* y *paciente*.
- [GÉNERO : MASCULINO/FEMENINO/NEUTRO]. Indica el género gramatical del objeto referenciado, pudiendo ser *masculino*, *femenino* o *neutro*. Se trata al masculino como género no marcado, prevaleciendo en cualquier referencia de múltiples objetos de distinto género.
- [NÚMERO : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el objeto referenciado, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos.

- [PERSONA : PRIMERA/SEGUNDA/TERCERA]. Distingue entre la implicación del objeto referenciado en el acto comunicativo, denominando *primera persona* al emisor, *segunda persona* al receptor y *tercera persona* a cualquier otro objeto no categorizable en los anteriores.

- [SIGNO : ADITIVO/SUSTRACTIVO]. Nota el signo con el que ha de marcarse una propiedad o evento, siendo *aditivo* si se posee o se considera, respectivamente, y *sustractivo* de no ser así.

Entre los rasgos de una lexía y sus homónimos en las lexías que la modifican, según los requisitos para su validez y el rasgo resultante tras la modificación, se observa una de las siguientes tres relaciones:

1. *Relación de atribución*. Se reemplaza el rasgo original por el rasgo modificador, salvo que este último no se encuentre presente en las lexías modificadoras. Se considera inválida si existe disparidad entre los rasgos de las lexías modificadoras, no pudiéndose determinar el rasgo resultante.

2. *Relación de concordancia*. Se precisa de equivalencia entre su rasgo original y el rasgo modificador para su validez, manteniendo el rasgo original en el caso de no encontrarse presente en las lexías modificadoras.

3. *Relación de descarte*. Se ignora el rasgo modificador, manteniendo su rasgo original e, incluso, ausentándolo de no estar presente en la lexía. Al depender exclusivamente de la misma, esta relación es siempre considerada válida.

Existen lexías que poseen diversas acepciones de su conjunto $R$ de rasgos gramaticales donde, exclusivamente, se muestra una variación en los valores de algunos de los rasgos gramaticales presentes. En estos casos, se decide agrupar los valores aceptados para un mismo rasgo gramatical, formando *rasgos polivalentes*. Su representación se realiza como se muestra a continuación:

$r = [\text{RASGO} : \text{VALOR } 1 + \text{VALOR } 2 + \cdots]$

Cuando se forman relaciones entre dos rasgos donde uno o ambos resultan ser polivalentes, puede darse la situación donde alguno de los valores aceptados resulta en una relación inválida. En dichos casos, se procede con la eliminación de los valores que causan tales relaciones inválidas.

## 6.2. Las lexías semánticas

Las lexías semánticas son aquellas que dotan al enunciado en el que estas se encuentran de un cierto valor semántico e intrínseco de las lexías pertenecientes a esta clase. Sin embargo, las lexías semánticas carecen de otra función que la aportación de dicho valor, por lo que dependen de otras clases para la determinación de los objetos referenciados, al igual que para la formación de cualquier jerarquía avanzada entre las lexías. El valor semántico de las mismas se traslada en forma de rasgos gramaticales completamente dependientes del significado, por lo que el análisis de los rasgos presentes en estas se limita exclusivamente a rasgos generales y puramente gramaticales.

Adicionalmente, en el ámbito morfológico del estudio lingüístico del español, destacan por ser la única clase de lexías generalmente divisibles en unidades de menor tamaño, denominadas como *monemas*. Estas unidades pueden considerarse, según su aportación a la lexía formada, como:

1. *Monemas adhesivos*. Carecen de propósito semántico o gramatical, pues se se sitúan entre otros monemas para su adecuada unión.

2. *Monemas gramaticales*. Aportan los rasgos gramaticales presentes en la lexía, además de indicar la subclase de lexía semántica en cuestión. Son monemas gramaticales todo afijo sin valor semántico, además de los morfemas flexivos.

3. *Monemas semánticos.* Dotan a la lexía del significado semántico propio de esta clase de lexías. Debido a la elevada variedad de monemas semánticos existentes, estos vuelven a las lexías semánticas el principal constituyente del repertorio léxico del español. Se incluyen aquí los lexemas, además de ciertos afijos.

Por último, las lexías semánticas, dependiendo de la localización y la función que estas desempeñan dentro de las distintas partes que componen un enunciado, se pueden dividir en cuatro subclases, mencionadas respectivamente con los rasgos generalmente presentes en las lexías de cada subclase:

- Adjetivos. Aporta una propiedad individual sobre un objeto referenciado, o compara dicha propiedad del mismo con la de otros objetos.
  - [GÉNERO : MASCULINO/FEMENINO]. Indica el género gramatical del objeto que recibe la propiedad, pudiendo ser *masculino* o *femenino*. En el caso de los adjetivos de doble terminación, este rasgo resulta ser polivalente, pudiendo aportar dicha propiedad a objetos masculinos y femeninos.
  - [GRADO : POSITIVO/COMPARATIVO/SUPERLATIVO]. Marca la relación resultante tras la aportación de la propiedad, siendo este *positivo*, si se limita a aportar la propiedad, *comparativo*, si compara el nivel de dicha propiedad presente entre el objeto referenciado y otro objeto, o *superlativo*, si expresa el nivel máximo de la propiedad respecto a un ámbito de relevancia.
  - [NÚMERO : SINGULAR/PLURAL]. Especifica el número del objeto que recibe la propiedad, distinguiéndose entre *singular*, si se aporta a un único objeto, y *plural*, si se aporta a múltiples objetos.
  - [PROPIEDAD : CALIFICATIVO/RELACIONAL]. Distingue entre adjetivos *calificativos*, que aportan propiedades intrínsecas del objeto, y adjetivos *relaciones*, que aportan propiedades que emergen de relaciones entre el objeto y su entorno.
  - [SIGNO : ADITIVO/SUSTRACTIVO]. Nota el signo con el que se aporta la propiedad, siendo *aditivo* si indica la presencia de dicha propiedad, o *sustractivo* si indica la ausencia de esta.

- □ [TERMINACIÓN : ÚNICA/DOBLE]. Puede ser de terminación *única*, si el adjetivo no varía según su género, o de terminación *doble*, si este posee versiones distintas para el género masculino y femenino.

■ *Adverbios*. Expresa el grado, el modo u otras características circunstanciales de una propiedad o un evento.

- □ [LÉXICO : SÍ/NO]. Indica si el valor semántico del adverbio en cuestión *sí* es léxico, si existe sin dependencia alguna, o *no* es léxico, si este requiere de un entorno para la consolidación de su valor.
- □ [SIGNO : ADITIVO/SUSTRACTIVO]. Nota el signo del valor expresado por el adverbio, siendo *aditivo* si indica la presencia de dicho valor, o *sustractivo* si indica la ausencia de este.

■ *Sustantivos*. Referencia a un objeto cualquiera, pudiendo este ser una instancia específica o el concepto general del mismo.

- □ [ARGUMENTAL : SÍ/NO]. Marca la capacidad del sustantivo de recibir argumentos, *sí* siendo argumental si posee dicha capacidad, y *no* siéndolo si carece de ella. Los sustantivos argumentales mayoritariamente proceden de la deverbalización de una cláusula, donde se preservan los argumentos de esta.
- □ [CANTIDAD : CONTINUO/DISCRETO]. Limita las formas en las que distintas cantidades del objeto referenciado por el sustantivo pueden ser nombradas, siendo *continuo* si se presenta en cantidades no numerales o dependientes de un sistema unitario, o *discreto* si se puede cuantizar en cantidades numerales enteras.
- □ [GÉNERO : MASCULINO/FEMENINO]. Indica el género gramatical del objeto referenciado, pudiendo ser *masculino* o *femenino*. Al no existir sustantivos de género neutro, este valor se considera inválido.
- □ [NÚMERO : SINGULAR/PLURAL]. Denota el número del objeto referenciado, distinguiéndose entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos.

- [DETERMINADO : SÍ/NO]. Especifica si se referencia a una instancia específica de un objeto, donde *sí* está determinado, o si se referencia al concepto del mismo, donde *no* está determinado. Un sustantivo solo puede estar determinado sin depender de ningún determinante cuando este referencia a un objeto con nombre propio.

■ *Verbos.* Especifica el evento de relevancia en una cláusula, ya sea la cláusula principal del enunciado o una cláusula subordinada dentro de esta.

- [GÉNERO : MASCULINO/FEMENINO]. En el caso de algunos tiempos no conjugables, indica el género gramatical del verbo, pudiendo ser *masculino* o *femenino*.
- [MODO : INDICATIVO/SUBJUNTIVO/IMPERATIVO]. Marca el modo del evento especificado por el verbo, siendo este *indicativo*, si indica la realización de una acción, *subjuntivo*, al relegar el modo del evento a cláusulas que lo contengan, o *imperativo*, si demanda la realización del evento a un agente externo. No se encuentra presente en los tiempos no conjugables.
- [NÚMERO : SINGULAR/PLURAL]. En el caso de algunos tiempos no conjugables, denota el número gramatical del verbo, distinguiéndose entre *singular* y *plural*.
- [PERSONA : PRIMERA/SEGUNDA/TERCERA]. Distingue entre la implicación del agente (o del paciente, si se encuentra en una cláusula pasiva) en el acto comunicativo, denominando *primera persona* al emisor, *segunda persona* al receptor y *tercera persona* a cualquier otro objeto no categorizable en los anteriores.
- [TIEMPO : $\cdots$]. Según la distancia temporal al momento de realización del evento en cuestión, los posibles tiempos de un verbo son: *condicional compuesto, condicional simple, futuro compuesto, futuro simple, gerundio compuesto, gerundio simple, infinitivo compuesto, infinitivo simple, participio compuesto, participio simple, presente, pretérito anterior, pretérito imperfecto, pretérito perfecto compuesto, pretérito perfecto simple* y *pretérito pluscuamperfecto*.

- [VARIANTE : -RA/-SE]. Indica la terminación morfológica del verbo en los tiempos verbales donde se presenta la opción.

## 6.3. Las lexías determinantes

Las lexías determinantes son las encargadas de determinar a objetos referenciados por otras lexías y, en casos, referenciar a los objetos mismos. De esta manera, las lexías determinantes son capaces de actuar de forma independiente para la referenciación de objetos. Sin embargo, se distinguen de las lexías semánticas al carecer de un valor semántico e intrínseco, por lo que dependen de estas para la referenciación de objetos por medio otro que la *deíxis*, proceso por el que se referencian a objetos previamente conocidos o presentes en el contexto lingüístico.

Hay ocasiones en las que más de una lexía determinante puede estar determinando a una lexía semántica o, incluso, formando una única referencia a un objeto sin depender de lexía semántica alguna. En estos casos, se toma una de las lexías como principal (la lexía semántica o la lexía determinante con menor índice de prioridad, respectivamente), y se consideran al resto como modificadores de esta. Como las lexías determinantes, además de determinar, aportan ciertas propiedades con valor gramatical, como la cantidad y la posesión, este mecanismo permite la especificación de más de una propiedad en la misma referencia. Hay que tener en cuenta, sin embargo, que dos lexías determinantes con el mismo índice de prioridad no se pueden encontrar modificando a la misma lexía o como referencia independiente.

Además, las lexías determinantes se distinguen morfológicamente de las lexías semánticas en su estricta indivisibilidad, pues se constituyen exclusivamente de una única unidad morfológica. Estas, por otro lado, se encuentran cuantizadas y limitadas a un conjunto reducido e inmutable, por lo que existe un número significativamente menor de lexías determinantes.

Por último, las lexías determinantes, según la propiedad característica de esta, se pueden agrupar en diversas subclases, todas ellas acompañadas por su correspondiente índice de prioridad:

- *Artículos (1)*. Indica el conocimiento previo sobre el objeto referenciado dentro del entorno en cuestión.
  - [DEFINIDO : SÍ/NO]. Especifica la presencia de previo conocimiento sobre el objeto, *sí* estando definido si es conocido respecto al entorno, o *no* estándolo de no ser así.
  - [GÉNERO : MASCULINO/FEMENINO/NEUTRO]. Indica el género gramatical del objeto referenciado, pudiendo ser *masculino*, *femenino* o *neutro*. En el caso de referenciar a un objeto de género neutro, este se ha de hallar como lexía principal de la referencia.
  - [NÚMERO : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el objeto referenciado, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos. Este es siempre singular, en el caso de referenciar a un objeto de género neutro.
- *Cuantificadores (2)*. Especifica la cantidad relativa en la que se presencia el objeto referenciado.
  - [GÉNERO : MASCULINO/FEMENINO]. Indica el género gramatical del objeto referenciado, pudiendo ser *masculino*, *femenino* o *neutro*. En el caso de referenciar a un objeto de género neutro, este se ha de hallar como lexía principal de la referencia.
  - [NÚMERO : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el objeto referenciado, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos. Este es siempre singular, en el caso de referenciar a un objeto de género neutro.
  - [PREDETERMINANTE : SÍ/NO]. Limita su función como posible predeterminante, especificando si *sí* puede actuar como uno, o si *no* es posible. Los únicos predeterminantes en el español son las variaciones de género y número de «todo».
  - [RANGO : PARCIAL/UNIVERSAL]. Si el cuantificador denota a un conjunto nulo o absoluto respecto a un entorno o marco de referencia, este se considera de rango

*universal*. Sin embargo, si dicho conjunto contiene una cantidad no absoluta de elementos, o es imprecisa, se considera de rango *parcial*.

- *Demostrativos (1)*. Aporta la distancia del emisor o receptor al objeto referenciado en niveles discretos, actuando además como una forma de señalamiento.
  - [DISTANCIA : CORTA/MEDIA/LARGA]. Especifica el nivel discreto de distancia en el que se encuentra el objeto referenciado respecto a un punto de referencia, pudiendo ser distancia *corta*, *media* o *larga*.
  - [GÉNERO : MASCULINO/FEMENINO/NEUTRO]. Indica el género gramatical del objeto referenciado, pudiendo ser *masculino*, *femenino* o *neutro*. En el caso de referenciar a un objeto de género neutro, este se ha de hallar como lexía principal de la referencia.
  - [NÚMERO : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el objeto referenciado, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos. Este es siempre singular, en el caso de referenciar a un objeto de género neutro.
- *Interrogativos (1)*. Referencia a cierta propiedad de un objeto por la que se pregunta, o introduce una cláusula subordinada que sustituye al valor de dicha propiedad.
  - [GÉNERO : MASCULINO/FEMENINO/NEUTRO]. Indica el género gramatical del objeto referenciado, pudiendo ser *masculino*, *femenino* o *neutro*. Los interrogativos frecuentemente carecen de género marcado morfológicamente, por lo que se suele considerar un rasgo polivalente en la práctica.
  - [NÚMERO : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el objeto referenciado, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos.
- *Numerales (2)*. Aporta la cantidad, el orden o la división, de forma exacta y numeral, en la que se presencia el objeto referenciado.
  - [FORMA NUMÉRICA : CARDINAL/ORDINAL/PARTITIVO]. Nota la forma numérica del

número expresado, siendo *cardinal* si expresa una cantidad, *ordinal* si expresa orden o *partitivo* si expresa una subdivisión de la unidad.

- □ [GÉNERO : MASCULINO/FEMENINO]. Indica el género gramatical del objeto referenciado inferior en la relación formada, pudiendo ser *masculino* o *femenino*. El género se encuentra, en la mayoría de casos, como rasgo polivalente, pues no se encuentra distinguido morfológicamente.
- □ [NÚMERO : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el objeto referenciado, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos.

■ *Posesivos (3)*. Denota el poseedor del objeto referenciado o a la base de la relación presentada en este, respecto al entorno en cuestión.

- □ [GÉNERO : MASCULINO/FEMENINO]. Indica el género gramatical del objeto referenciado inferior en la relación formada, pudiendo ser *masculino* o *femenino*. Solo se encuentra presente cuando actúa como modificador posterior.
- □ [NÚMERO : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el objeto referenciado inferior en la relación formada, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos.
- □ [NÚMERO (SUPERIOR) : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el poseedor u objeto referenciado superior en la relación formada, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos.
- □ [PERSONA (SUPERIOR) : PRIMERA/SEGUNDA/TERCERA]. Distingue entre la implicación del poseedor u objeto referenciado superior en la relación formada en el acto comunicativo, denominando *primera persona* al emisor, *segunda persona* al receptor y *tercera persona* a cualquier otro objeto no categorizable en los anteriores.

■ *Pronombres personales (1)*. Referencia a una entidad perteneciente al entorno en cuestión.

- □ [GÉNERO : MASCULINO/FEMENINO/NEUTRO]. Indica el género gramatical del objeto referenciado, pudiendo ser *masculino*, *femenino* o *neutro*. En el caso de referenciar a un objeto de género neutro, este se ha de hallar como lexía principal de la referencia.
- □ [NÚMERO : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el objeto referenciado, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos. Este es siempre singular, en el caso de referenciar a un objeto de género neutro.
- □ [PERSONA : PRIMERA/SEGUNDA/TERCERA]. Distingue entre la implicación del objeto referenciado en el acto comunicativo, denominando *primera persona* al emisor, *segunda persona* al receptor y *tercera persona* a cualquier otro objeto no categorizable en los anteriores.

■ *Relativos (1)*. Presenta un valor deíctico e introduce una cláusula subordinada, sustituyendo al objeto referenciado dentro de ella.

- □ [GÉNERO : MASCULINO/FEMENINO/NEUTRO]. Indica el género gramatical del objeto referenciado, pudiendo ser *masculino*, *femenino* o *neutro*. La mayoría de relativos carecen de distinción morfológica entre géneros, por lo que el rasgo frecuentemente se considera polivalente en la práctica.
- □ [NÚMERO : SINGULAR/PLURAL]. Marca la cantidad en la que se presencia el objeto referenciado, distinguiendo entre *singular*, si se referencia un único objeto, y *plural*, si se referencian múltiples objetos. Al igual que el género, este suele considerarse polivalente en la práctica.

## 6.4. Las lexías gramaticales

Las lexías gramaticales son las encargadas de aglutinar otras lexías en estructuras más complejas que las que ellas mismas pueden formar. Se distinguen del resto, particularmente, en que no aportan ningún valor o propiedad de ningún tipo a otras lexías, a la cláusula o a la totalidad del enunciado. Sin embargo, poseen la capacidad de agregar pausas textuales,

formar coordinaciones entre lexías y cláusulas de la misma clase, adaptar lexías para la modificación de otras y jerarquizar dicha modificación de lexías.

Estas, según su función principal, se agrupan en las siguientes subclases:

- *Conjunciones*. Según la estructura que marque, forma jerarquías entre lexías o coordina varias lexías o cláusulas de la misma clase en una única.
  - [ESTRUCTURA : COORDINANTE/SUBORDINANTE]. Indica el tipo de estructura marcada por la lexía, siendo *coordinante* si aglutina varias lexías o cláusulas en una única, o *subordinante* si introduce una cláusula subordinada.
  - [RELACIÓN : ADVERSATIVA/COPULATIVA/DISYUNTIVA]. En el caso de tratarse de una conjunción coordinante, especifica la relación entre las lexías o las cláusulas coordinadas, siendo *adversativa* si las relaciona como elementos opuestos, *copulativa* si las agrupa en una única unidad o *disyuntiva* si propone la selección exclusiva de una de ellas.
- *Preposiciones*. Modifican el rango de funciones realizables por una lexía, adaptándolas para la modificación de otras lexías o para su adecuada situación como argumentos en su cláusula. Las preposiciones carecen de rasgos gramaticales generalmente presentes.
- *Símbolos*. Según su tipo, extienden coordinaciones para incluir a un mayor número de lexías o cláusulas, rompen o terminan divisiones estructurales en el enunciado o añaden pausas textuales. Los símbolos carecen de rasgos gramaticales generalmente presentes.

# 7. Nivel sintáctico

Tras el nivel léxico-morfológico, el nivel sintáctico es el siguiente nivel a tratar en el modelo, pues es el encargado de describir las siguientes estructuras que resultan de la modificación, coordinación y jerarquización de lexías, mencionando con exactitud su composición, sus requisitos de formación y el orden de los elementos dentro de dichas estructuras. Para ello, se introduce como unidad principal *el sintagma*.

Por otro lado, las lexías y los sintagmas resultantes precisan de alguna estructura adicional que los contenga, formando el enunciado en cuestión, pues este carece del rigor que las estructuras sintagmáticas presentan en su formación, orden y requisitos. Para ello, se introduce, además, la unidad de *la cláusula*, ya previamente mencionada.

## 7.1. Notación de Backus-Naur

Para poder mantener un moderado rigor matemático a la hora de describir las estructuras resultantes de las operaciones previamente descritas, se utiliza una notación proveniente de la rama científica de la computación y la algoritmia, conocida como la notación de Backus-Naur (BNF, o *Backus-Naur form*, en inglés) [Backus & Naur, 1959], que se basa en las reglas generativistas de Noam Chomsky. Esta es considerable como un metalenguaje formalizado para expresar gramáticas libres de contexto, y es frecuentemente utilizada para la descripción de lenguajes de programación de forma teórica.

La unidad básica en esta notación, *la regla*, se define especificando su composición de forma ordenada. Para ello, a cada regla se le asigna una expresión que define dicha estructura. En su forma elemental, esta regla se define de la siguiente forma:

$$\langle \text{REGLA} \rangle := \textit{expresión}$$

Las expresiones de las reglas se componen de una de las siguientes formaciones posibles:

- *literal*. La primera expresión elemental es el literal, consistiendo este en un conjunto de caracteres en cursiva delimitado por espacios, que forman una lexía específica. Esto es utilizable para definir ciertas reglas a partir de lexías específicas, o en casos donde el rango de lexías posibles para dicha regla sea lo suficientemente reducido como para su enumeración.

- ⟨REGLA⟩. La última expresión elemental es la sustitución de otra regla dentro de ella, permitiendo la creación de estructuras recursivas y jerarquías de reglas, operación fundamental para la definición de una gramática libre de contexto.

- *expresión expresión*. La siguiente expresión elemental es la concatenación incondicional de otras, mediante su situación delimitada entre ellas por espacios. Al ser una operación recursiva, su aplicación para más de dos expresiones es divisible en parejas compatibles con esta operación.

- *expresión* | *expresión*. La barra actúa como operador de selección, por lo que una expresión de esta forma equivale a exclusivamente una de las dos expresiones nombradas en la estructura. Esto permite la creación de reglas que pueden concordar con dos o más estructuras distintas.

- [*expresión*]. Al rodear la expresión por corchetes, se indica que la presencia de la misma es opcional, equivaliendo de forma selectiva tanto a esa expresión como a una expresión vacía. Esta operación es la que permite la omisión de ciertas partes dentro de algunas reglas. Si esta contiene un subíndice posterior a ella, $[\textit{expresión}]_1$, se presente limitar la omisión de la misma, garantizando la presencia de, al menos, una de las expresiones marcadas con el mismo subíndice.

- {*expresión*}. Al delimitar la expresión por llaves, se marca la opcionalidad de su presencia. Sin embargo, a diferencia de la anterior, la expresión puede aparecer múltiples veces, es decir, esta se encuentra cero o más veces. Si esta contiene una cruz posterior a ella, {*expresión*}+, se prohíbe su omisión, apareciendo una o más veces necesariamente. En cambio, si esta contiene un subíndice posterior a ella, $\{\textit{expresión}\}_1$, se pretende limitar la omisión de la misma, garantizando la presencia de, al menos, una de las expresiones marcadas con el mismo subíndice.

- (*expresión*). Al cerrar la expresión por paréntesis, se consigue delimitar a esta, con el fin de marcar la extensión de los operadores cercanos, pues estos se limitan a incluir a las expresiones inmediatamente anteriores y posteriores, salvo que se utilice esta misma operación como agrupación.

Es importante tener en cuenta durante la lectura de las descripciones de las reglas en el modelo que las reglas para todas las subclases de lexías presentes en cada clase se encuentran predefinidas con los siguientes nombres: ⟨ADJETIVO⟩, ⟨ADVERBIO⟩, ⟨SUSTANTIVO⟩, ⟨VERBO⟩, ⟨ARTÍCULO⟩, ⟨CUANTIFICADOR⟩, ⟨DEMOSTRATIVO⟩, ⟨INTERROGATIVO⟩, ⟨NUMERAL⟩, ⟨POSESIVO⟩, ⟨PRONOMBRE PERSONAL⟩, ⟨RELATIVO⟩, ⟨CONJUNCIÓN⟩ y ⟨PREPOSICIÓN⟩.

Por último, debido a la naturaleza combinatoria de las estructuras definidas con esta notación y por la inexactitud de la lengua natural, se puede dar el caso donde el mismo enunciado sea analizable resultando en más de una estructura distinta. Como todas las estructuras resultantes podrían ser válidas desde un punto de vista sintáctico, se precisa de una comprensión semántica para el descarte de las estructuras semánticamente inválidas. La búsqueda de una solución se encuentra fuera del alcance del trabajo de investigación, sin embargo, se proporciona una aproximación funcional y suficientemente precisa en la implementación del modelo.

## 7.2. El sintagma

El sintagma es la estructura resultante de la agrupación de una lexía semántica junto a toda otra lexía o sintagma que la modifique. De esta forma, la mayoría de la jerarquía resultante de la aplicación de modificadores proviene de la existencia de sintagmas que los incluya, frecuentemente de forma recursiva. Como se muestra en la posterior definición de las estructuras sintagmáticas, con el fin de simplificar las reglas que las definen, se tratan a las lexías carentes de modificadores como sintagmas que contienen una única lexía.

El sintagma posee un núcleo, compuesto de una lexía semántica u otro sintagma, sobre el que se aplican el resto de lexías y sintagmas presentes en el mismo. Además, las lexías

dentro de las estructuras resultantes han de seguir un orden específico para su formación, el cual se encuentra establecido y limitado entre sus diversas opciones.

Según la subclase de lexía semántica que actúe de núcleo en el sintagma, estos son clasificables en las siguientes clases, acompañados con su definición completa en notación Backus-Naur:

- *Sintagma adjetival.* Poseen un adjetivo como núcleo, y su estructura es la siguiente:

$$\langle \text{S.A.} \rangle ::= \{\langle \text{PREPOSICIÓN} \rangle\}\ [\langle \text{S.R.} \rangle]\ (\langle \text{ADJETIVO} \rangle \mid \langle \text{POSESIVO} \rangle)\ [\langle \text{S.N.} \rangle \mid \langle \text{C.} \rangle]$$

- *Sintagma adverbial.* Poseen un adverbio como núcleo, y su estructura es la siguiente:

$$\langle \text{S.R.} \rangle ::= \{\langle \text{PREPOSICIÓN} \rangle\}\ [\langle \text{ADVERBIO} \rangle]\ \langle \text{ADVERBIO} \rangle$$

- *Sintagma nominal.* Poseen un sustantivo como núcleo, y su estructura es la siguiente:

$$\begin{aligned}\langle \text{S.N.} \rangle ::= {} & \{\langle \text{PREPOSICIÓN} \rangle\}\ [\langle \text{PREDETERMINANTE} \rangle] \\ & (\{\langle \text{DETERMINANTE} \rangle\}\ [\langle \text{S.A.} \rangle]\ [\langle \text{SUSTANTIVO} \rangle]_1\ \{\langle \text{S.A.} \rangle \mid \langle \text{S.N.} \rangle \mid \langle \text{C.} \rangle\}_1 \\ & \mid \{\langle \text{PREPOSICIÓN} \rangle\}\ [\langle \text{PREDETERMINANTE} \rangle]\ \langle \text{PRONOMBRE PERSONAL} \rangle) \\ \langle \text{PREDETERMINANTE} \rangle ::= {} & \textit{todo} \mid \textit{toda} \mid \textit{todos} \mid \textit{todas} \\ \langle \text{DETERMINANTE} \rangle ::= {} & \langle \text{ARTÍCULO} \rangle \mid \langle \text{CUANTIFICADOR} \rangle \mid \langle \text{DEMOSTRATIVO} \rangle \mid \langle \text{INTERROGATIVO} \rangle \\ & \mid \langle \text{NUMERAL} \rangle \mid \langle \text{POSESIVO} \rangle\end{aligned}$$

- *Sintagma verbal.* Poseen un verbo como núcleo, y su estructura es la siguiente:

$$\begin{aligned}\langle \text{S.V.} \rangle ::= {} & ([\langle \text{DEÍCTICO PACIENTE} \rangle]\ [\langle \text{DEÍCTICO OBJETO} \rangle]\ \langle \text{V. (T. CONJUGABLE)} \rangle) \\ & \mid (\langle \text{V. (T. NO CONJUGABLE)} \rangle\ [\langle \text{DEÍCTICO PACIENTE} \rangle]\ [\langle \text{DEÍCTICO OBJETO} \rangle]) \\ \langle \text{DEÍCTICO PACIENTE} \rangle ::= {} & \textit{me} \mid \textit{te} \mid \textit{le} \mid \textit{nos} \mid \textit{os} \mid \textit{les} \mid \textit{se} \\ \langle \text{DEÍCTICO OBJETO} \rangle ::= {} & \textit{me} \mid \textit{te} \mid \textit{lo} \mid \textit{la} \mid \textit{nos} \mid \textit{os} \mid \textit{los} \mid \textit{las}\end{aligned}$$

## 7.3. La cláusula

La cláusula es la unidad que agrupa múltiples sintagmas, entre ellos necesariamente un sintagma verbal que actúe como núcleo, y forma una estructura completa y de orden libre. Esta tiene la función de expresar un evento, identificado con el núcleo, opcionalmente aportando información contextual a partir del resto de sintagmas presentes en la misma cláusula. El enunciado, unidad máxima analizada por el modelo, es considerado también una cláusula, y es la que engloba al resto de lexías del mismo.

En contraste con el funcionamiento del núcleo en los sintagmas, en la cláusula, el núcleo no es modificado por el resto de lexías o sintagmas presentes dentro de la cláusula, sino que la agrupación de ambas presentan una unidad sintáctica y semántica inmutable. La razón, pues, de la existencia de un núcleo es para justificar la imposibilidad de su omisión, a diferencia de el resto de lexías o sintagmas pertenecientes a la misma cláusula.

Las lexías o sintagmas que no actúan como núcleo dentro de una cláusula son denominados argumentos, y se pueden clasificar según su función en las siguientes:

- *Agente*. Referencia al objeto que realiza el evento, al que se le denomina como agente.
- *Circunstancial*. Añade contexto circunstancial y semántico sobre el estado del entorno de realización del evento descrito por el núcleo de la cláusula.
- *Modificador*. Modifica de forma indirecta a uno de los tres objetos referenciables dentro de la cláusula: el agente, el objeto o el paciente. No se considera parte de la lexía o sintagma que describe a estos objetos referenciables, pues pueden existir otros argumentos de la cláusula entre ambos. Aún así, se exige la misma concordancia que en los modificadores de una lexía.
- *Objeto*. Referencia al objeto de relevancia en el evento, frecuentemente inanimado, sobre el que no necesariamente recae el mismo evento. A este, en el contexto de la cláusula, se le denomina como objeto.
- *Paciente*. Referencia al objeto sobre el que se realiza o recae el evento en cuestión, razón por la que es denominado como paciente.

Según su posicionamiento relativo a la estructura del enunciado y su propia estructura interna, las cláusulas son clasificables en las siguientes clases y subclases, explicadas brevemente:

- *Cláusulas principales*. Engloban al enunciado, conteniendo la totalidad de las lexías presentes en él dentro de ella. Se clasifican según la clase de enunciado resultante, que depende de la intención general de este en el acto comunicativo:
  - *Cláusulas declarativas*. Poseen la intención de declarar la realización de un evento, ya sea en el pasado, en el presente o como posibilidad o certeza en el futuro, o como una realidad atemporal.
  - *Cláusulas exclamativas*. Enfatizan alguna propiedad conocida de uno de los argumentos presentes en la cláusula.
  - *Cláusulas interrogativas*. Cuestionan la realización del evento en cuestión, o preguntan sobre las propiedades de uno de los argumentos presentes en la cláusula.
  - *Cláusulas imperativas*. Demandan la realización del mismo evento a un acusado que realiza la función de agente dentro de la cláusula.
- *Cláusulas subordinadas*. Están contenidas dentro de la cláusula principal o de otra cláusula subordinada, y presentan un evento secundario ocurrente dentro de estas. Además, son frecuentemente introducidas por ciertas lexías exclusivas para este propósito, o por verbos en tiempos no conjugables. Se clasifican en diversas clases, dependiendo de la forma en la que son introducidas:
  - *Cláusulas adjuntas temporales*. Se introducen por una conjunción subordinante, y expresa una relación temporal entre la cláusula que la contiene y esta, actuando siempre como argumento de la cláusula. El verbo que describe el evento de esta, en ocasiones, puede encontrarse en un tiempo perteneciente al modo subjuntivo.
  - *Cláusulas argumentales*. Siempre son introducidas por las conjunciones subordinantes «que» o «si», y actúan como un sustantivo o un sintagma nominal.

En las cláusulas argumentales comenzadas por la conjunción «que», el verbo que describe el evento de esta se encuentra necesariamente en un tiempo perteneciente al modo subjuntivo.

- *Cláusulas de gerundio o participio*. Comienzan por un verbo en los tiempos no conjugables de gerundio o de participio, y actúan exclusivamente como argumento de una cláusula.
- *Cláusulas de infinitivo*. Son introducidas, como su nombre indica, por un verbo en los tiempos no conjugables de infinitivo, actuando como un sustantivo o un sintagma nominal.
- *Cláusulas relativas*. Se introducen por los mismos interrogativos que introducen a las cláusulas interrogativas, denominados en este caso como relativos, sin embargo, actúan como el argumento por el que se pregunta dentro de esta.

## 7.4. Coordinación

La coordinación es el proceso por el cual se consigue la aglutinación horizontal de una o más lexías, sintagmas o cláusulas en una única unidad, compartiendo ciertas propiedades y rasgos. Este proceso es frecuentemente utilizado para expresar las relaciones de agrupación, oposición o selección exclusiva con las unidades que componen la unidad resultante. Se requiere para la realización del mismo que las unidades coordinadas sean semejantes, es decir, se traten del mismo tipo de unidad y pertenezcan a la misma clase y subclase.

Según la relación que se forma, identificable a partir de la conjunción coordinante en uso, se clasifican en las siguientes clases:

- *Coordinación adversativa*. Opone a las unidades pertenecientes a esta.
- *Coordinación copulativa*. Agrupa varias unidades, actuando como su unión.
- *Coordinación disyuntiva*. Propone la selección exclusiva de una de las unidades.

## 8. Nivel semántico

Finalmente, el nivel semántico es el último nivel detallado en el modelo, al ser el encargado de analizar semánticamente el significado del enunciado mediante la asignación de roles y la clasificación de los argumentos pertenecientes a las cláusulas del mismo. Además, cumple la función de descartar ciertas estructuras jerárquicas considerables como inválidas. Para llevarlo a cabo, se introduce *el rol semántico* como herramienta de marcado.

El rol semántico es una marca realizada a los argumentos de una cláusula que especifica su función respecto al evento en cuestión, y el procedimiento para su asignación depende particularmente del significado semántico del verbo que marca el evento en cuestión. Por lo tanto, para su asignación, se ha de utilizar una base de datos de verbos, emparejados con las combinaciones posibles de roles semánticos que se pueden dar, como, en el caso de la implementación posteriormente mencionada del modelo, *AncoraNet* [Antònia Martí et al., 2007].

Los roles semánticos reconocidos en este modelo, basándose en los presentes en la base de datos utilizada para su validez y concordancia, son los siguientes, acompañados cada uno con su abreviación técnica utilizada generalmente en otras investigaciones:

- *Lugar* (LOC).
- *Extensión* (EXT).
- *Conectiva del discurso* (DIS).
- *Propósito general* (ADV).
- *Marca de negación* (NEG).
- *Verbo modal* (MOD).
- *Causa* (CAU).
- *Tiempo* (TMP).
- *Propósito* (PNC).

- *Modo* (MNR).
- *Dirección* (DIR).
- *Predicación secundaria* (PRD).

Sin embargo, debido a la subjetividad involucrada en el proceso de análisis y asignación de roles semánticos, se ha decidido prestar mayor importancia a la clasificación por funciones anteriormente descrita, pues, a diferencia de los roles semánticos, son fácilmente asignables por medios algorítmicos y rigurosos. De esta manera, se justifica la falta de contenido en el apartado del nivel semántico del modelo, siendo considerablemente menos extenso que el resto.

Aun así, en la posterior implementación del modelo, se redefinirá el apartado correspondiente al nivel semántico para incluir, también, la asignación de funciones a los argumentos presentes dentro de las cláusulas que componen al enunciado.

# Parte III. Implementación del modelo

El modelo anterior proporciona una visión determinista de la gramática del español, que es de grata utilidad para la implementación de sistemas y herramientas de análisis de enunciados. Por lo tanto, se propone a continuación una implementación tipo de tal sistema, explicando con detalle los algoritmos, las estructuras de datos y los formatos utilizados para ello.

Al tratarse exclusivamente de un posible método para su implementación, no se garantiza que el diseño de esta implementación sea óptimo, ya sea en los ámbitos de utilización de memoria, complejidad de tiempo o carga de datos. Asimismo, la transpilación de lo presentado en este apartado a código, presente en el repositorio de GitHub «https://github.com/segfaultdev/grammar_es/» junto con una copia de esta memoria, puede contener errores de funcionamiento no diagnosticados, por lo que se ruega discreción a la hora de realizar su lectura.

La implementación propuesta se divide en los mismos niveles que el modelo utilizado y descrito en este trabajo de investigación, considerándolos a continuación (son utilizados ciertos anglicismos para la denominación de las partes del sistema, tomados de la jerga computacional y repetidos a lo largo de este apartado, pues carecen de equivalente perfecto en el español):

- *Nivel léxico-morfológico.* Mediante un lexicalizador o «lexer», se lexicalizan los conjuntos de caracteres que componen el enunciado de entrada, resultando en una o más descomposiciones del mismo en lexías pertenecientes al listado de lexías utilizado.
- *Nivel sintáctico.* Un analizador o «parser» se encarga de transformar cada una de estas descomposiciones que recibe como entrada en una o más jerarquías de lexías, sintagmas y cláusulas para cada descomposición ingresada.
- *Nivel semántico.* Un etiquetador o «tagger» le asigna, para cada estructura obtenida, las funciones desenvueltas por los argumentos de las cláusulas contenidas a estos

mismos, utilizando para ello un listado de las funciones correspondientes a cada verbo, y devolviendo tras ello toda combinación válida de funciones.

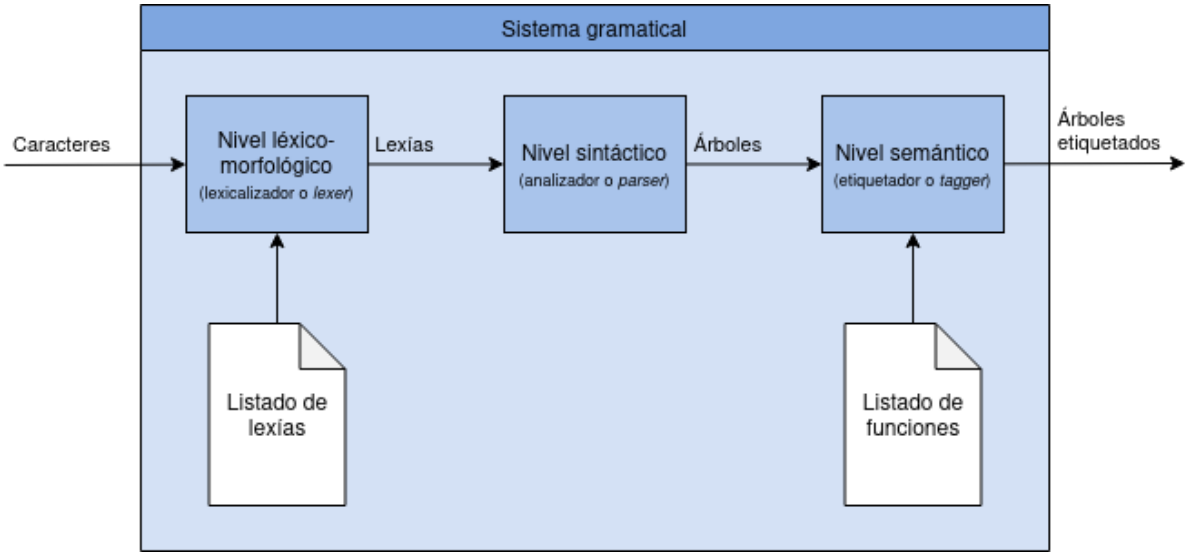


Fig. 1: Esquema del sistema gramatical (preliminar).

Como resulta aparente en la organización del sistema implementado, cada uno de los niveles va aumentando considerablemente el número de combinaciones aceptadas como válidas. Esto puede llegar a ser un considerable problema, especialmente cuando el enunciado recibido posee múltiples interpretaciones o es excesivamente extenso. Además, algunas de las acepciones pueden resultar más o menos cercanas al significado y la intención con los que originalmente se concibió el enunciado, por lo que resultaría ideal la reducción de estas a un número realisticamente manejable. Se proponen, como posibles soluciones, las dos siguientes medidas, ambas implementadas en el sistema tipo propuesto:

- *Filtrado de árboles.* Se modifica el «parser» para que, en ciertos estados intermedios de la creación de las diferentes acepciones de posibles estructuras de árbol resultantes, utilice una serie de filtros preestablecidos para la eliminación de árboles que, por razones relacionadas con la forma estructural o la concordancia, son considerables inválidos. Este proceso consigue, en la práctica, reducir el número de árboles en varios órdenes de magnitud.

- *Puntuado de árboles.* Al final de la cadena de niveles, se añade un puntuador de árboles, cuyo propósito es utilizar ciertas reglas seleccionadas manualmente para poder discernir acepciones según la posibilidad aproximada de coincidir con la intención original. Sin embargo, los criterios utilizados para la asignación de puntuaciones son empíricos y subjetivos, por lo que, para evitar errores procedentes de la selección de dichos criterios, no se descarta ningún árbol, limitándose a la reordenación por puntuación.

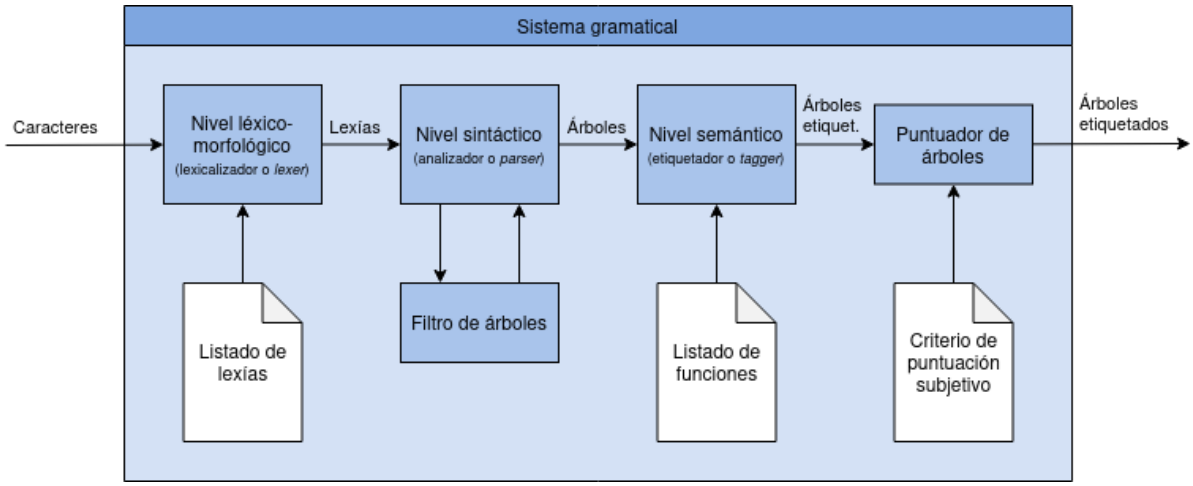


Fig. 2: Esquema del sistema gramatical (final).

# 9. Herramientas en uso

En la implementación tipo propuesta, se han utilizado diversas herramientas específicas, con el fin de facilitar su diseño y programación. Debido a la existencia de múltiples herramientas que permiten la creación de sistemas gramaticales, se ha decidido utilizar las herramientas enumeradas a continuación:

- *C23 (GCC 13.2)*. El lenguaje de programación C, aun minimalista en la funcionalidad incluida por defecto, tanto en la librería estándar como en la sintaxis del lenguaje, otorga una elevada eficiencia, al ser un lenguaje compilado y de bajo nivel. Desde su introducción en 1972, ha marcado un antes y un después en el ámbito de la computación, aunque la versión utilizada para la implementación, C23, se encuentra en la etapa final previa a su publicación, por lo que es considerable un lenguaje moderno. Este ha sido utilizado para la programación del sistema gramatical, situado en el «backend» de la plataforma de análisis resultante. Además, es el lenguaje de preferencia de los autores intelectuales del trabajo de investigación.

- *HTML5* y *CSS2*. Estos dos lenguajes de marcado y estilizado, respectivamente, han hecho la construcción de una interfaz gráfica atractiva y accesible al público general mediante el dominio web asignado, por lo que han sido utilizados para la programación del «frontend» de la plataforma de análisis resultante.

- *HTMX (v1.19.10)*. La librería HTMX permite la creación de interfaces gráficas interactivas y ligeras, accesibles por medio de una página web. En este caso, ha sido utilizada para la conexión entre la interfaz accesible al público, escrita en HTML5 y CSS2, y el sistema gramatical, escrito en C23, permitiendo la solicitud de análisis de enunciados en tiempo real.

# 10. Nivel léxico-morfológico

El nivel léxico-morfológico es el encargado de producir como resultado, dado un enunciado en forma de cadena de caracteres, todas las acepciones posibles de divisiones de dichos caracteres en diferentes lexías, según un listado de lexías.

En este caso, se ha utilizado la tabla de frecuencias de elementos gramaticales disponible en la página web del Corpus del español del siglo XXI [Real Academia Española [RAE], 2023], pues consigue aportar un extenso y completo listado de las lexías presentes en el español, incluyendo 2.554.444 lexías distintas. Además, la misma tabla de frecuencias aporta un análisis morfológico exhaustivo para cada una de las lexías presentes, clasificándolas en las subclases existentes y mencionadas en el modelo, y otorga valores para la gran mayoría de rasgos correspondientes a cada una de las subclases. De esta forma, se tiene una única fuente centralizada y verificada de lexías, que luego es procesada y filtrada hasta convertirla al formato requerido en el sistema gramatical.

## 10.1. Rasgos gramaticales

Se crea una estructura, denominada `trait_t`, que alberga las propiedades de un rasgo gramatical específico, conteniendo su nombre (`name`), un conjunto con los valores que posee (`values`) y la cantidad de dichos valores (`value_count`), siendo esta más de uno exclusivamente en el caso de los rasgos polivalentes. Tanto el nombre como los valores son guardados en el formato UTF-8.

```
typedef struct trait_t trait_t;

struct trait_t {
    const char8_t *name;

    const char8_t *values;
    size_t value_count;
};
```

## 10.2. Lexías

Se crea una estructura, denominada `word_t`, que almacena las propiedades de una lexía específico, conteniendo su nombre (`name`), un conjunto con los rasgos gramaticales que posee (`traits`) y la cantidad de dichos rasgos gramaticales (`trait_count`). El nombre de la lexía corresponde con su descomposición por caracteres, donde todas las letras alfabéticas se encuentran en minúsculas, salvo la primera que, si se trata de un sustantivo propio, se puede encontrar en mayúscula.

Además, se crea un conjunto global de todas las lexías presentes en el listado que se encuentre en uso (`words`), acompañado de la cantidad de dichas lexías (`word_count`).

```
typedef struct word_t word_t;

struct word_t {
    const char8_t *name;

    const trait_t *traits;
    size_t trait_count;
};

/*  Global  */
word_t *words = NULL;
size_t word_count = 0;
```

## 10.3. Árboles de prefijos

Los árboles de prefijos son estructuras de datos utilizadas para almacenar un conjunto de cadenas de caracteres que poseen ciertos prefijos en común entre sí. Esto se consigue mediante utilización de nodos que, salvo el nodo padre, almacenan un caracter, y que poseen una arista saliente para cada caracter que podría continuar al anterior. Esto significa que cada ruta que se puede trazar desde el nodo padre hasta un nodo terminal, nodo que se encuentra marcado como el final de una cadena de caracteres, representa a una cadena de caracteres distinta.

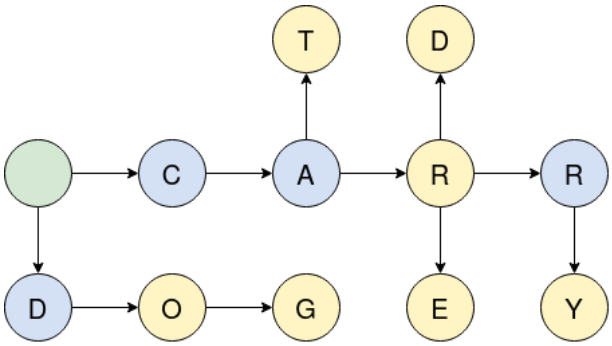


Fig. 3: Ejemplo de árbol de prefijos, para las cadenas «car», «card», «care», «carry», «cat», «do» y «dog». El nodo verde es el nodo padre y los nodos amarillos son nodos terminales.

Para la simplificación de las estructuras almacenadas en cada nodo, es frecuente trazar aristas únicamente a uno de los nodos y unir el resto a este mediante una cadena de aristas, conectando cada uno con su siguiente. Esto permite que cada nodo tenga, como máximo, dos aristas: una que apunta hacia un nodo hijo, y otra que apunta hacia un nodo hermano.

La razón por la que los árboles de prefijos son utilizados en el sistema gramatical presentado es porque permiten la división de caracteres en cadenas en complejidad de tiempo lineal, es decir, que depende exclusivamente del número de caracteres que se solicite dividir, y no del número de lexías presentes en el listado de lexías utilizado para la creación del árbol de lexías. Por otro lado, facilitan considerablemente la compleja tarea de explorar todas las combinaciones de divisiones posibles, pues se vuelve una tarea relativamente simple al realizarse recursivamente sobre una estructura de árbol. La única desventaja aparente es el elevado tiempo necesitado para la creación del mismo, aunque este se puede realizar una única vez, almacenándolo para evitar su repetida generación.

Estos presentan, sin embargo, un significativo problema al utilizarlos para la división de caracteres en lexías: existen cadenas de caracteres que representan a más de una lexía, por lo que se tiene que encontrar alguna forma de discernir estas lexías entre sí. Para ello, se propone almacenar en cada nodo terminal todas las lexías que resultan de la cadena formada por la ruta entre dicho nodo y el nodo padre.

Se crea una estructura, denominada `prefix_t`, que almacena las propiedades de una nodo específico, conteniendo el caracter que representa (`data`), un conjunto con los índices de las lexías que son representadas por la ruta del nodo padre al nodo en cuestión (`indexes`),

la cantidad de dichos índices (`index_count`), el índice del nodo hermano (`sibling`), o cualquier número negativo de no existir tal nodo, y el índice del nodo hijo (`child`), o cualquier número negativo de no existir tal nodo. Los caracteres en las rutas para cada lexía han de corresponder con los presentes en el nombre de la misma lexía.

Además, se crea un conjunto global de todos los nodos presentes en el árbol de prefijos formado (`prefixes`), acompañado de la cantidad de dichos nodos (`prefix_count`).

```
typedef struct prefix_t prefix_t;

struct prefix_t {
    char8_t data;

    size_t *indexes;
    size_t index_count;

    ssize_t sibling, child;
};

/* Global */
prefix_t *prefixes = NULL;
prefix_t prefix_count = 0;
```

# 11. Nivel sintáctico

El nivel sintáctico es el encargado, para cada división resultante del lexicalizador, de hallar las estructuras de jerarquías y coordinaciones de lexías, sintagmas y cláusulas, denominadas árboles, que se pueden formar a partir de las lexías recibidas como entrada al analizador.

Se ha utilizado con este fin las formas en las que se encuentran definidos los sintagmas y las cláusulas, utilizando la notación de Backus-Naur. Esta, aunque normalmente es utilizada para definir gramáticas no ambiguas, ha demostrado a lo largo del trabajo de investigación su eficacia para la definición de cualquier gramática libre de contexto. Al ser así, el analizador únicamente ha de utilizar estas definiciones para intentar encajar cada lexía dentro del árbol resultante. En el caso de las cláusulas, de hecho, no es necesario ninguna forma que las defina, pues los argumentos y el núcleo presentes en ella pueden encontrarse, en principio, en cualquier orden.

## 11.1. Analizador de descenso recursivo

Un analizador de descenso recursivo (*Recursive descent parser*, en inglés) es un tipo de analizador gramatical utilizado para transformar entradas lineales, en este caso compuestas de lexías, en árboles, conforme a una gramática libre de contexto. Generalmente, estos analizadores se programan de tal forma que actúan como analizadores «zurdos», es decir, que priorizan los árboles cuyo desbalance se encuentra a la izquierda. Sin embargo, al tratarse el modelo propuesto en el trabajo de investigación de una gramática ambigua, este es modificado para poder producir más de un árbol como resultado.

Estos analizadores utilizan como base rutinas separadas para cada forma posible, que llaman recursivamente a otras, integrando progresivamente unidades procedentes de la entrada al árbol resultante y retrocediendo cuando una forma es determinada inválida a lo largo de este proceso, hasta poder continuar con en análisis. En el caso de la versión modificada, sin embargo, se procede con dicho retroceso incluso si la generación del árbol es exitosa, con el fin de poder hallar todos los árboles que cumplen con las formas preestablecidas en el modelo.

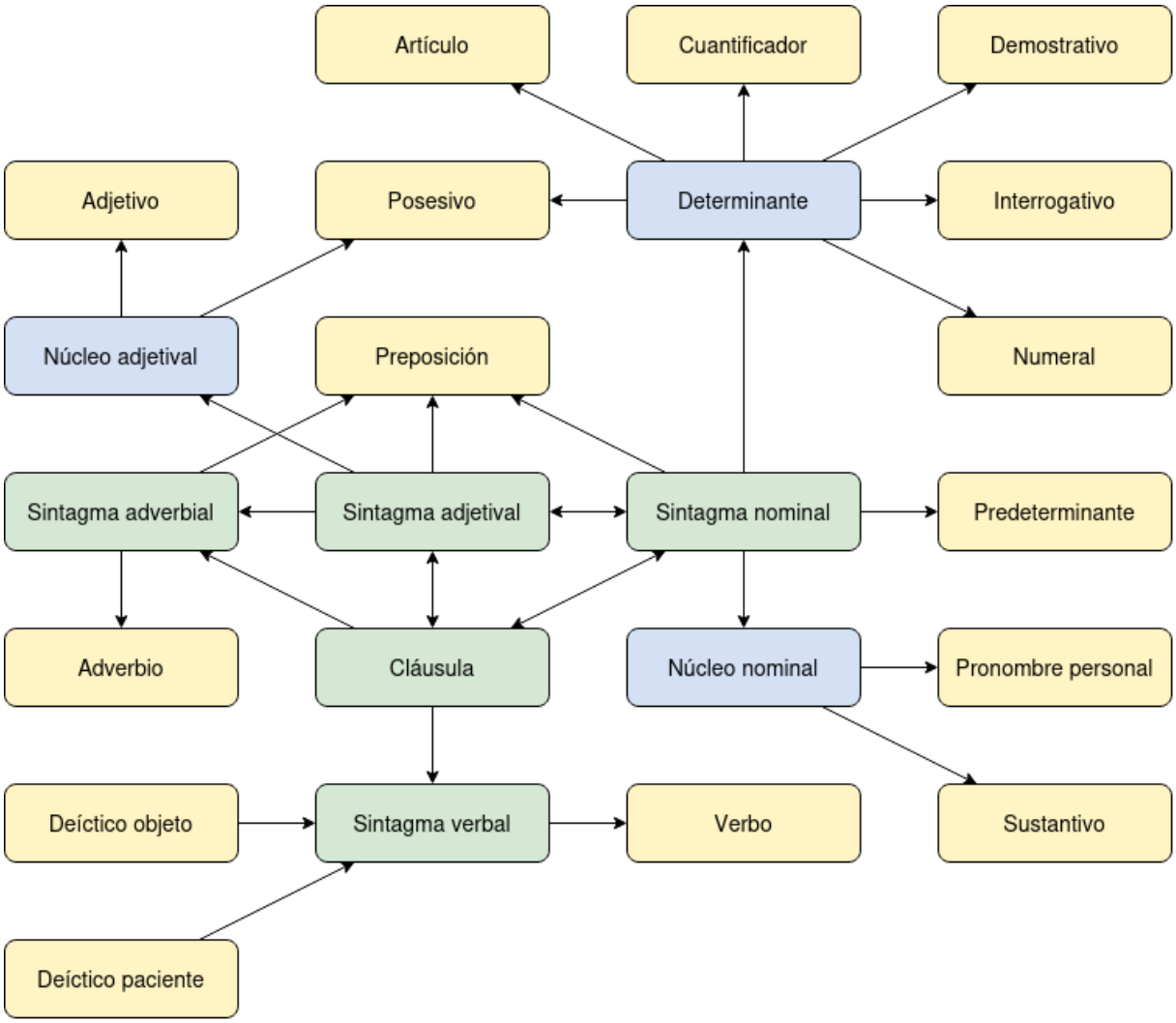


Fig. 4: Diagrama de recursión en el sistema gramatical. Los vértices verdes son rutinas externas, los vértices azules son rutinas internas y los vértices amarillos son subclases de lexías.

La razón por la que se ha decidido utilizar un analizador de descenso recursivo es por su alta eficiencia combinada con la facilidad presentada para su implementación, pues esta consiste en una traducción de la notación Backus-Naur que describe la gramática libre de contexto a llamadas recursivas de rutinas, que son creadas para cada vértice en el diagrama de recursión del sistema gramatical.

## 11.2. Filtro de árboles

A lo largo del proceso de generación de todas las acepciones de árboles en el analizador, se pueden llegar a producir cantidades excesivas de árboles, siendo muchos de ellos inválidos tras un rápido filtrado. Por lo tanto, tras cada rutina del diagrama, se introduce

un filtro de árboles. Este consigue reducir en varios órdenes de magnitud el número de árboles resultantes, acelerando masivamente el sistema gramatical, reduciendo su consumo de memoria y permitiendo, pues, el análisis de enunciados de mayor longitud.

El analizador, por sí solo, aumenta en cada recursión el número de árboles de forma recursiva, razón por la que es necesario este filtro. Esto ocurre debido a la intencional simplicidad del mismo, pues, para conseguir un sistema breve y autoexplicativo, se ha decidido utilizar únicamente la clase y subclase de las lexías analizadas para la creación preliminar de los árboles, previos al filtrado y, en el nivel semántico, etiquetado.

Las medidas utilizadas para el filtrado eficaz de árboles se encuentran enumeradas a continuación, junto a una breve descripción de las mismas:

- *Discordancia de rasgos*. En muchos árboles, algunos rasgos que requieren de concordancia, como el género, el número y la persona, pueden no poseer valores equivalentes, resultando en relaciones inválidas, por lo que estos árboles son descartados.

- *Repetición de árboles*. Existe más de una ruta que puede resultar diversos árboles con la misma forma, incluyendo únicamente variaciones ligeras en los rasgos gramaticales o, incluso, no variando de ninguna manera, por lo que pueden ser reducidos a un único árbol.

- *Formas degeneradas*. En las formas donde varias reglas son opcionales, puede darse el caso donde se produzca una omisión o una inclusión excesiva de reglas presentes en la forma, resultando en formas que no cumplen en su totalidad las reglas definidas en el modelo, por lo que los árboles correspondientes se eliminan.

# 12. Nivel semántico

El nivel semántico es el encargado, para cada acepción procedente del analizador del nivel sintáctico, de asignar las funciones de los argumentos dentro de las cláusulas y de algunos sintagmas, además de corregir y reducir algunos rasgos presentes en ciertas lexías y sintagmas.

El listado de verbos procedente del corpus *AnCora* [Antònia Martí et al., 2007] proporciona algo menos de 3.000 verbos distintos en español, junto a las distintas funciones que estos aceptan cuando se encuentran como núcleo de una cláusula. Además, se dispone de la forma necesaria de los argumentos para la asignación de cada función posible, junto a las funciones que un argumento puede desempeñar, extraído previamente a partir de los árboles generados.

Con todo estos recursos, en el sistema gramatical, se utilizan métodos de selección combinatorios y de «backtracking» para la asignación de estas funciones, asignándolas de cualquier forma arbitraria hasta llegar a un argumento al que no se le puede asignar ninguna función, o hasta completar la asignación de funciones, donde se retrocede para seguir comprobando el resto de asignaciones válidas.

De esta manera, para cada árbol originario del analizador de descenso recursivo, se crean varias copias, cada una con una asignación válida de funciones distinta. A su vez, si algún árbol procedente del analizador no posee ninguna asignación válida de funciones, es descartado. Finalmente, rasgos como la función, el rol semántico y otros rasgos polivalentes son finalmente separados en las distintas combinaciones de valores que pueden resultar.

## 13. Puntuador de árboles

Finalmente, previo a la salida de los árboles generados al exterior del sistema, estos son puntuados mediante diversos criterios, seleccionados manualmente para la minimización del error del sistema gramatical a la hora de analizar diversos enunciados. Tras ser puntuados, se ordenan según la puntuación, y se expulsan del sistema. La puntuación asignada corresponde, de manera aproximada, con la inversa de la probabilidad de su aparición de forma natural, asignando menores puntuaciones a árboles más naturales.

Es importante tener en cuenta, sin embargo, que este último paso no descarta ningún árbol, ni selecciona un subconjunto de estos árboles para su salida tras la ordenación por puntuación, sino que se limita a devolverlos en el orden generado. Se ha decidido realizar el proceso de esta manera para evitar el descarte de acepciones válidas que, por alguna imprecisión al calcular su puntuación, han sido valoradas con una puntuación más elevada que la esperada. De la misma forma, estos mismos errores pueden provocar la aparición de árboles inválidos con puntuaciones menores que las deseadas. Esto está destinado a ocurrir eventualmente, pues los criterios seleccionados se eligieron de forma empírica y subjetiva, pero por medio de la intensiva comprobación del sistema gramatical se ha pretendido mitigar lo mejor posible.

Cada criterio otorga una puntuación decimal, con el $i$-ésimo criterio asignando la puntuación $1 \leq x_i \leq +\infty$. La fórmula para calcular la puntuación total de un árbol, siendo $n$ el número de criterios presentes, es la siguiente:

$$X = \prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n$$

La cantidad de criterios utilizados en la implementación del modelo excede la cantidad descriptible en el trabajo de investigación, además de que su descripción se encuentra fuera del alcance del mismo, por lo que se proporcionan las ideas detrás de los dos criterios utilizados de mayor relevancia:

- *Frecuencia de acepciones de lexías homólogas y rasgos polivalentes*. Según la frecuencia registrada en el listado de lexías para cada acepción de las lexías homólogas presentes en el enunciado, se puede conseguir priorizar usos más frecuentes de ciertas lexías sobre usos escasamente registrados. Además, según los valores finalmente utilizados para rasgos considerados polivalentes tras el último nivel del sistema, ciertos árboles se pueden considerar más naturales que otros.

- *Horizontalidad de los modificadores de lexías*. Debido al considerable esfuerzo mental que un ser humano ha de realizar para preservar con exactitud información contextual del enunciado sin recurrir a las múltiples repeticiones del mismo, es considerablemente más frecuente ver cadenas de modificadores de lexías y sintagmas, donde cada uno modifica al anterior, que múltiples modificadores afectando a una única lexía o a un único sintagma. Por lo tanto, se puede utilizar como criterio de puntuación la longitud de los modificadores que han de ser recordados como contexto, es decir, que se encuentren sucedidos por otros que afecten a la misma unidad gramatical.

# Parte IV. Epílogo

## 14. Conclusiones

Con este trabajo de investigación se pretendía diseñar un modelo adaptado de la gramática española que se fundamentase en una visión más determinista y matemáticamente rigurosa, dejando a un lado cualquier interpretación del significado más allá del extraíble de forma objetiva mediante un análisis semántico donde se contraste las formas encontradas con las existentes en listados. Además, se buscaba que este modelo fuera lo más cercano con la sintaxis completa del español, minimizando de toda manera posible los casos donde este modelo no consiguiera definir de forma exacta la jerarquía y coordinación que se presencia dentro del enunciado analizado. Sin embargo, se tomó la natural decisión de limitar el alcance al enunciado, omitiendo cualquier análisis textual en el trabajo de investigación, pues escalaría considerablemente la complejidad del mismo, al ser un apartado de la gramática que, tradicionalmente, ha requerido de una mayor subjetividad a la hora de su análisis.

Por otro lado, se propuso como objetivo secundario la creación de una herramienta de uso fácil e inmediato para el análisis de estos mismos enunciados, u oraciones, de tal manera que fuera disponible de forma totalmente gratuita al público mediante una interfaz interactiva en una página web, además de construir un sistema gramatical interno y modular que funcionase como traducción directa del modelo descrito en el trabajo de investigación, generando estos análisis no solo para la herramienta web, sino como sistema implementable dentro de otros programas y servicios.

Estas dos metas principales del trabajo resultan ser considerablemente ambiciosas, pues son exclusivamente beneficiosas para la sociedad, permitiendo innovaciones en los ámbitos de educación, tecnología y legislación, entre otros. Por ende, este trabajo de investigación puede considerarse una sólida base para futuras expansiones e investigaciones sobre el tema de los modelos deterministas, proporcionando un mayor rigor y una mayor fiabilidad, en comparativa con la gran mayoría de aproximaciones actualmente disponibles al público, tanto libres como comerciales.

Sin embargo, durante la realización del trabajo de investigación, se ha realizado una reflexión sobre las consideraciones éticas a tener en cuenta, siendo los dos aspectos más problemáticos los siguientes:

- *La utilización de los datos del usuario para la mejora del sistema.* Esta opción fue planteada desde el comienzo del trabajo de investigación, con claros planes de su materialización y fragmentos de código escritos y verificados. A pesar de ello, previo a la redacción del sistema en el trabajo de investigación y la publicación del código fuente, se optó por desechar los mecanismos propuestos para ello, pues no se consideró que tuviera los beneficios suficientes como para considerarse una opción factible. De esta manera, la implementación del sistema carece, en su totalidad, de protocolos de telemetría y recolección de datos.

- *La integración del modelo y el sistema en modelos de inteligencia artificial con intenciones maliciosas.* Es cierto que, al publicar este trabajo de investigación, se puede dar el caso en el que se utilice la información recopilada en este trabajo con el fin de producir modelos de inteligencia artificial para fines como la suplantación de identidad. En este caso, se ha optado por marcar como irrelevante este aspecto, pues las implicaciones directas del trabajo no son lo suficientemente significativas como para justificar la toma de medidas al respecto, además de que este trabajo afecta positivamente a la rama de estudio de la lingüística.

Se ha de decir, con todo ello en cuenta, que se han logrado todos los objetivos propuestos con éxito, consiguiendo crear un modelo determinista y riguroso de los apartados de relevancia de la gramática del español. Esto se ha conseguido, adicionalmente, de una forma que resulte accesible y comprensible por tanto lingüistas, como programadores, como investigadores interesados. Por otro lado, se ha conseguido implementar una herramienta funcional y de fácil manejo, cuyo código fuente se encuentra disponible en la dirección «https://github.com/segfaultdev/grammar_es», con el fin de ser analizado, verificado y alojado en un servidor por el lector.

Como logro recalcable, además, la herramienta ha demostrado ser capaz de realizar análisis a tiempo real de oraciones de alrededor de 20 palabras sin error alguno y, en

general, muestra una tasa de funcionamiento considerablemente elevada, proporcionando el análisis correcto en las 3 primeras opciones en la gran mayoría de pruebas cerradas realizadas. Estas observaciones, aun subjetivas, se han obtenido mediante el análisis de un listado privado de enunciados de selección propia, cuya proveniencia se limita a fragmentos literarios del siglo XX. Se propone, como posible mejora, realizar un estudio cuantitativo y exhaustivo de la tasa de funcionamiento del sistema, pues no ha resultado posible dentro de los limitados recursos disponibles.

Aun así, se considera que el trabajo podría ser mejorado o extendido en ciertas áreas que, por falta de recursos, no se han podido tratar en el trabajo de investigación:

- *La realización de análisis textuales*, tanto de la estructura textual como de la semántica involucrada, realizando automáticamente la división por enunciados. Sería de relevancia, además, la realización de análisis de las propiedades textuales de *adecuación*, *coherencia* y *cohesión*, dentro de lo posible con un modelo determinista. Para ello, se tendría que expandir el modelo propuesto en este trabajo, junto al sistema algorítmico que lo implementa. La detección de vínculos deícticos en el texto, por otro lado, abriría puertas a su utilización en asistentes virtuales, proporcionando mejoras destacables.

- *La optimización del sistema gramatical* para el procesado de enunciados de mayor longitud sin demoras excesivamente extensas. En su actual versión, para garantizar el análisis válido y eficaz del enunciado, se ha optado por comprobar todas las descomposiciones léxicas posibles y, para cada una de ellas, todos los árboles posibles. Esto provoca que la herramienta requiera de un elevado tiempo de computación, además de altos requisitos de memoria. Es por ello por lo que se propone utilizar, en un futuro, algoritmos voraces para la aproximación óptima del análisis sin requerir de dichos requisitos temporales y de almacenamiento.

- *La búsqueda de mejores criterios de puntuación*, pues es una de las selectas áreas donde se carece de respaldo alguno, más allá de una breve experimentación paramétrica privada. Los dos posibles enfoques a llevar a cabo serían un enfoque empírico o de optimización, donde se ajustasen valores, ya sea de forma manual o automática, hasta

maximizar la tasa de funcionamiento; o un enfoque teórico, donde se expandiese el modelo propuesto para incluir criterios extrapolados directamente de la gramática española para la puntuación de enunciados.

En retrospectiva de lo dicho en esta sección y a lo largo del trabajo de investigación, se concluye dando por completo el trabajo de investigación, siendo este un «éxito rotundo», pues se han conseguido con creces todos los objetivos propuestos en este de forma detallada y respaldada con diversas fuentes involucradas en el campo de la lingüística, especialmente pertenecientes a las corrientes generativistas.

## Referencias

[Antònia Martí et al., 2007] Antònia Martí, M., Taulé, M., Bertran, M., & Màrquez, L. (2007). AnCora: Multilingual and Multilevel Annotated Corpora.

[Asociación de Academias de la Lengua Española [ASALE], Real Academia Española [RAE], 2009] Asociación de Academias de la Lengua Española [ASALE], Real Academia Española [RAE] (2009). *Nueva gramática de la lengua española*. Espasa-Calpe.

[Backus & Naur, 1959] Backus, J. & Naur, P. (1959). The syntax and the semantics of the proposed international algebraic language of the zurich acm-gamm conference.

[Instituto Cervantes, 2022] Instituto Cervantes (2022). *El español: una lengua viva. Informe 2022*. Technical report.

[Jackendoff, 1977] Jackendoff, R. (1977). X-bar syntax: A study of phrase structure.

[Real Academia Española [RAE], 2023] Real Academia Española [RAE] (2023). Corpus del Español del Siglo XXI.

[Zagona, 2006] Zagona, K. (2006). *Sintaxis generativa del español*.